Problem 1: Support Vector Machine (35 points)

Requirements:

1) Draw a line as the decision boundary that optimizes the above formulation.
2) Explain in short how you get that line, but you do not need to show a detailed proof.
3) Write a pair of $\mathbf{w}$, b which can define that decision boundary.
4) Circle the support vectors.

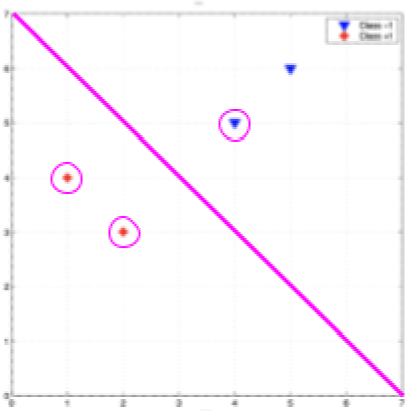
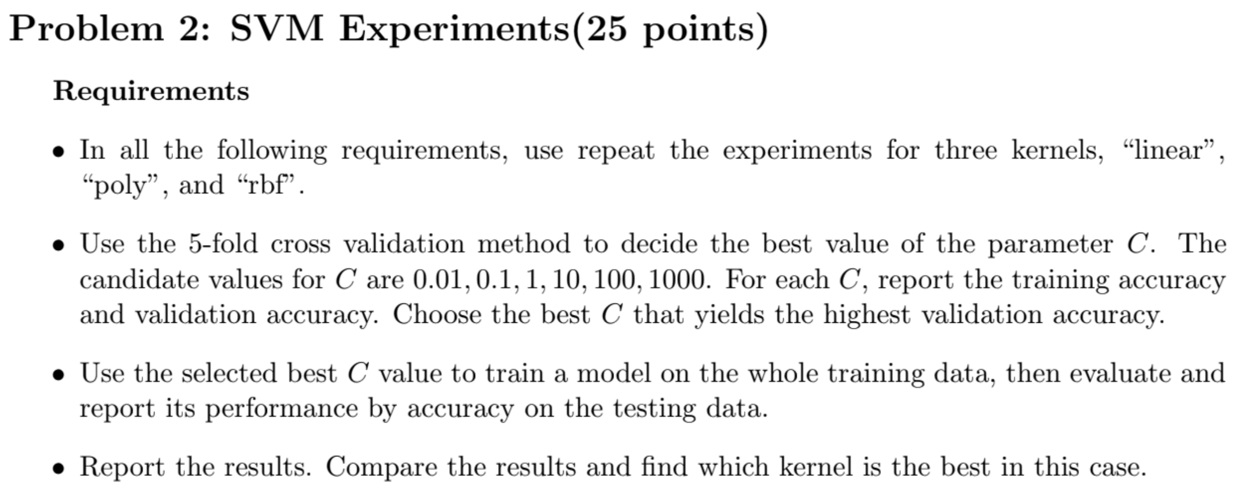


Figure 1: SVM

1. Line drawn in Magenta
2. That line creates an equal boundary between the closest points on both sides. That distance is $\sqrt{2}$ units.
3. $5w_1 + 2w_2 + b = 0 \Rightarrow \left(w_1, w_2, b\right) = (1, 1,\ -7)$
4. Circles drawn in Magenta



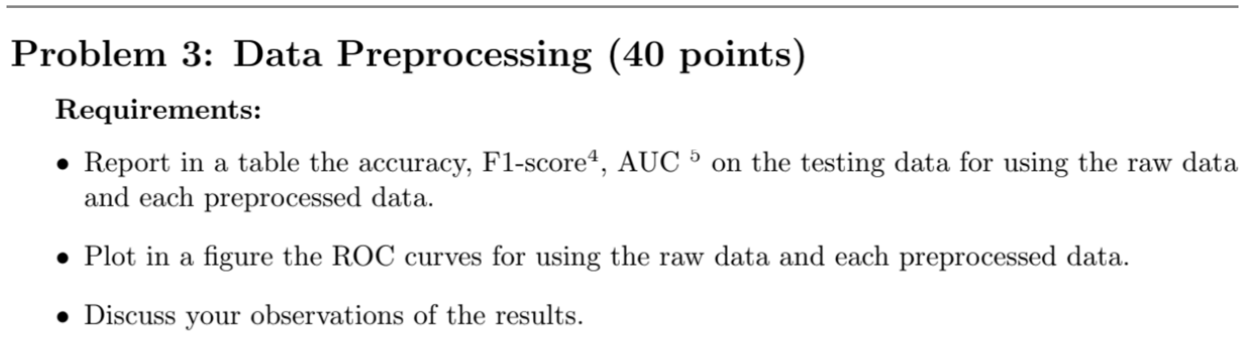
Problem 2: SVM Experiments(25 points)

Requirements

- In all the following requirements, use repeat the experiments for three kernels, “linear”, “poly”, and “rbf”.
- Use the 5-fold cross validation method to decide the best value of the parameter C. The candidate values for C are $0.01, 0.1, 1, 10, 100, 1000$. For each C, report the training accuracy and validation accuracy. Choose the best C that yields the highest validation accuracy.
- Use the selected best C value to train a model on the whole training data, then evaluate and report its performance by accuracy on the testing data.
- Report the results. Compare the results and find which kernel is the best in this case.

Kernel	Linear	Poly	RBF
Best C	100	10	10
Best C-V Accuracy	0.513	0.513	0.447
Test Set Accuracy	0.690	0.862	0.931

According to these results, the RBF kernel has the highest test set accuracy, despite having lower C-V Accuracy than the other two tests. For this particular case, RBF is the best kernel to use.



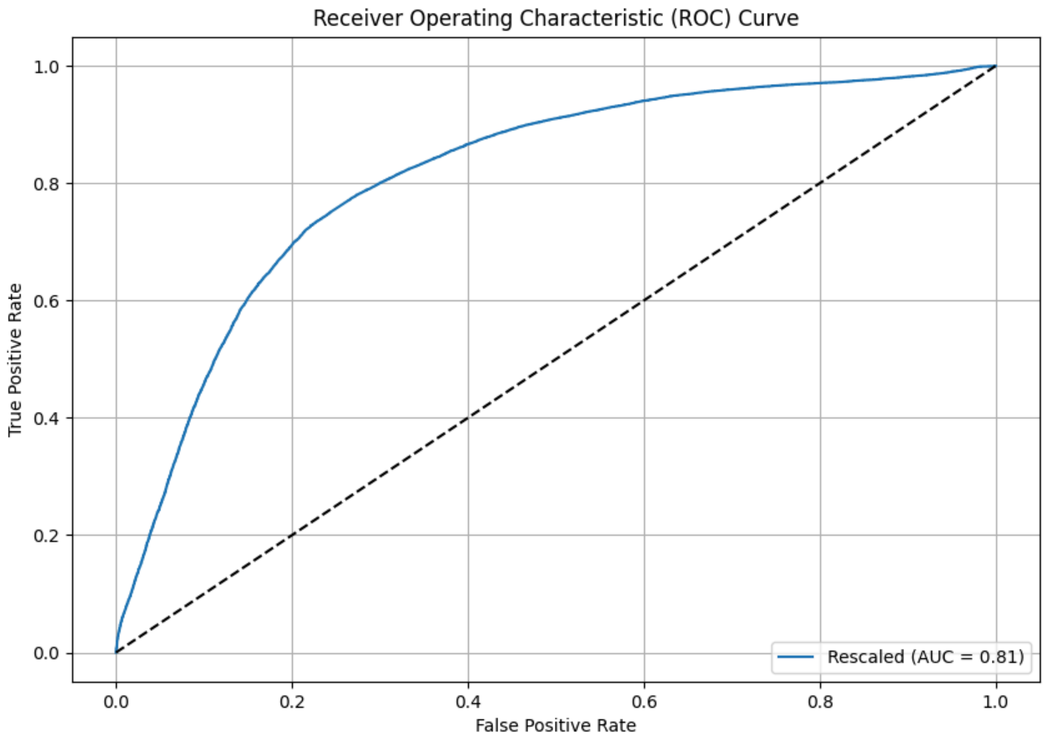
Problem 3: Data Preprocessing (40 points)

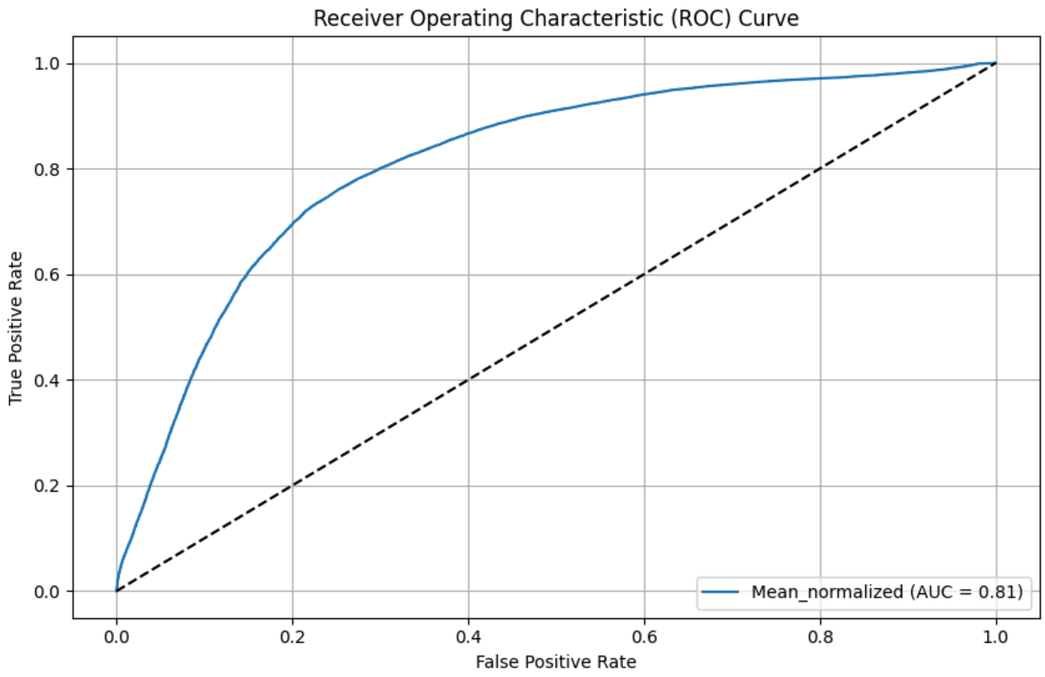
Requirements:

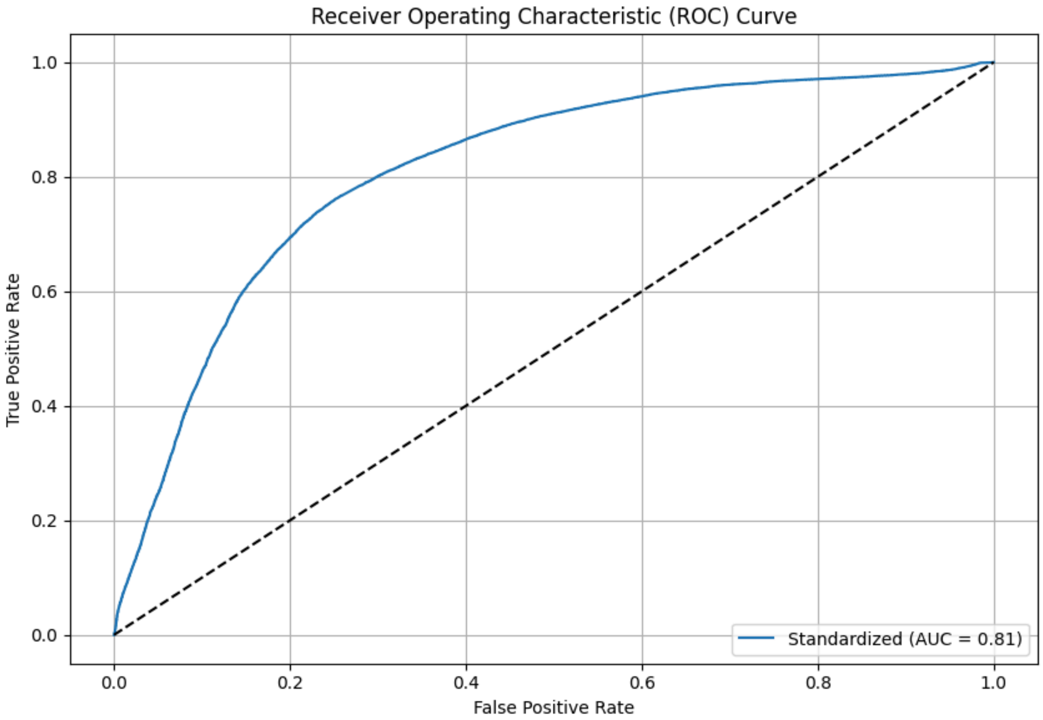
- Report in a table the accuracy, F1-score[4], AUC [5] on the testing data for using the raw data and each preprocessed data.
- Plot in a figure the ROC curves for using the raw data and each preprocessed data.
- Discuss your observations of the results.

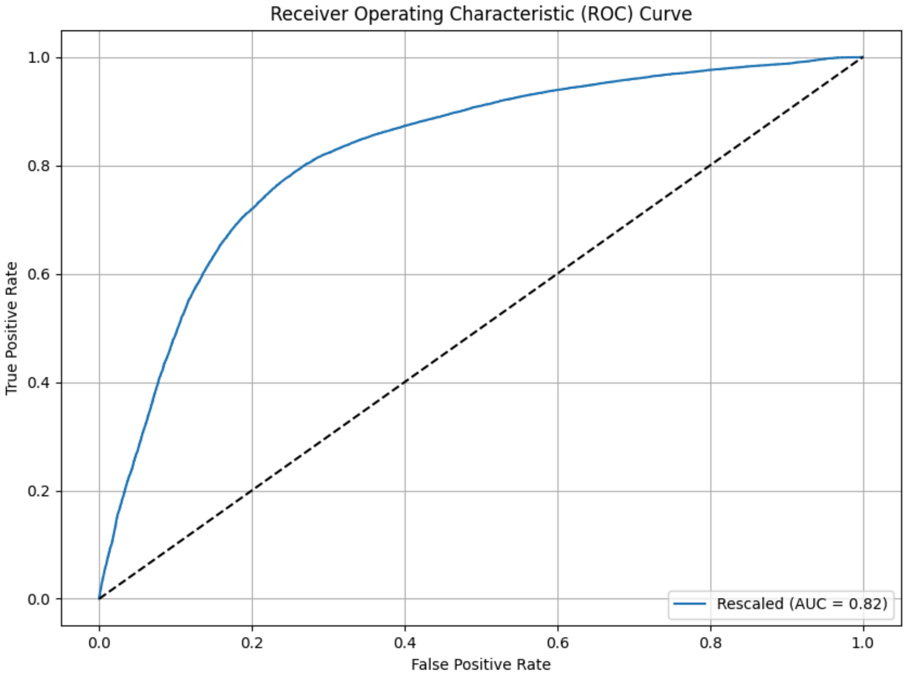
Method	Rescaled	Mean Normalized	Standardized
Test Accuracy	0.7521	0.7520	0.7508
Test F1 Score	0.7537	0.7538	0.7540
Test AUC	0.8129	0.8129	0.8128

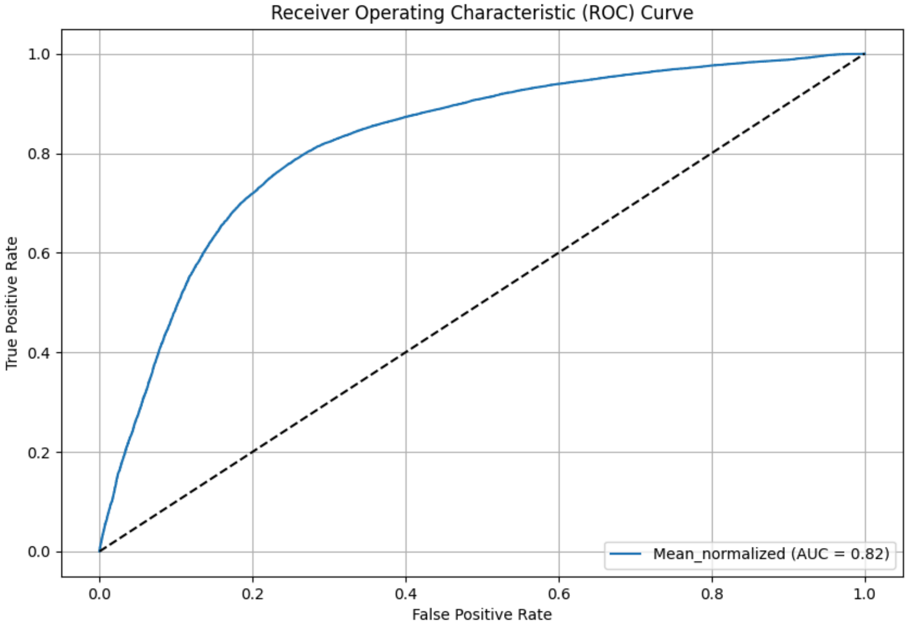
Each of these have very similar results, demonstrating that any of them are valid choices for preprocessing. Graphs on next page


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Rescaled (AUC = 0.81)


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean_normalized (AUC = 0.81)


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Standardized (AUC = 0.81)


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Rescaled (AUC = 0.82)


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Mean_normalized (AUC = 0.82)